Estimation of Rényi Information Divergence via Pruned Minimal Spanning Trees[1]

Alfred Hero
Dept. of EECS,
The university of Michigan,
Ann Arbor, MI 48109-2122, USA
Email: hero@eecs.umich.edu

Olivier J.J. Michel
Laboratoire de Physique, URA-1325 CNRS,
École Normale Supérieure de Lyon,
46 allée d'Italie,
69364 Lyon Cedex 07, France
Email: omichel@physique.ens-lyon.fr

Abstract

In this paper we develop robust estimators of the Rényi information divergence (I-divergence) given a reference distribution and a random sample from an unknown distribution. Estimation is performed by constructing the minimal spanning tree (MST) passing through the random sample points and applying a change of measure which flattens the reference distribution. In a mixture model where the reference distribution is contaminated by an unknown noise distribution one can use these results to reject noise samples by implementing a greedy algorithm for pruning the k-longest branches of the MST, resulting in a tree called the k-MST. We illustrate this procedure in the context of discrimination and robust clustering for a planar mixture model.

1. Introduction

Let $\mathrm{X}_n = \{x_1, x_2, \ldots, x_n\}$ denote a sample of i.i.d. data points in R^d having unknown Lebesgue multivariate density $f(x_i)$ supported on $[0,1]^d$. Define the order ν Rényi entropy of f [7]

$$H_\nu(f) = \frac{1}{1-\nu} \ln \int f^\nu(x) dx \qquad (1)$$

[1]This work was supported in part by AFOSR under MURI grant: F49620-97-0028.

and, for a dominating Lebesgue density f_o, the Rényi information divergence of f with respect to f_o

$$I_\nu(f, f_o) = \frac{1}{1-\nu} \ln \int \left(\frac{f(x)}{f_o(x)}\right)^\nu f_o(x) dx \qquad (2)$$

The quantity $I_\nu(f, f_o)$ is a special case of I-divergence which is called the Chernoff distance or the Renyi cross-entropy between f and f_o [1]. The I-divergence takes on its minimum value (equals zero) if and only if $f = f_o$ (a.e.). The Rényi information divergence $I_\nu(f, f_o)$ specializes to the Rényi entropy $H_\nu(f)$ when f_o is equal to a uniform density over $[0,1]^d$. Other special cases of interest are obtained for $\nu = \frac{1}{2}$ for which one obtains the log Hellinger distance squared

$$I_{\frac{1}{2}}(f, f_o) = \ln \left(\int \sqrt{f(x) f_o(x)} dx \right)^2$$

and for $\nu \to 1$ for which one obtains the Kullback-Liebler divergence

$$\lim_{\nu \to 1} I_\nu(f, f_o) = \int f_o(x) \ln \frac{f_o(x)}{f(x)} dx.$$

The problem of estimating the I-divergence arises in the very large class of density classification problems for clustering and pattern recognition [1, 3]. In these problems one applies a threshold test to an estimate of $I_\nu(f, f_o)$ in order to decide whether f is equal to f_o. I-divergence estimation also arises in image registration where the I-divergence can be directly related to mutual information between two images f and f_o [8]. For an overview of entropy and I-divergence estimation applications the reader can refer to [2] and [1].

In this paper we present a methodology for generating robust estimates of $I_\nu(f, f_o)$ for unknown f and arbitrary dominating density f_o. This methodology performs a non-linear transformation on the data sample X_n, producing a transformed data sample Y_n, and constructs a graph, called the k-minimal spanning tree (k-MST), on a minimal k-point subset $\mathrm{Y}_{n,k}$ of the transformed data. The k-MST is a graph which connects k out of n of the data points in a manner that minimizes the total length of the graph, where length is defined as the sum of the interconnection distances (called edges) raised to a user-specified power $\gamma \in (0, d)$. Here k plays the same role as the parameter α in the α-trimmed mean estimator of the ensemble mean of a sample: when there are outliers k can be selected to ensure outlier resistance. The log length of the k-MST gives an estimator of the Rényi information divergence which has the following properties

- when k is set to n one obtains the plain vanilla MST whose log length is an asymptotically unbiased and strongly consistent estimator of $I_\nu(f, f_o)$
- for $k \in \{2, 3, \ldots, n\}$ the k-MST estimator specifies a minimal set of k points $\mathrm{X}_{n,k}$
- the sequence of sets $\{\mathrm{X}_{n,k}\}_{k=2}^n$ specifies a quasi-linear ordering of $\{x_1, \ldots, x_n\}$ which provides an extension of the notion of a rank order statistic to multiple dimensions $d > 1$.
- the k-MST Rényi information divergence estimator is direct and does not require the difficult step of estimating the density f
- estimates of arbitrary orders $\nu \in (0, 1)$ of the Rényi information divergence are obtained by adjustment of the edge power exponent γ used in the construction of the k-MST
- the method extends straightforwardly to estimation of related I-divergence measures such as the α-structural entropy introduced by Havrda and Charvàt [**?**]

To illustrate our results we will show ROC curves for the MST estimates of Rényi information divergence and give an application to robust clustering for the case that f is a planar mixture density of the form

$$f = (1 - \epsilon) f_1 + \epsilon f_o, \tag{3}$$

where f_o is a known outlier density and f_1, $\epsilon \in [0, 1]$ are unknown.

2. MST's and Entropy Estimation

A spanning tree T through the sample X_n is a connected acyclic graph which passes through all the n points $\{x_i\}_i$ in the sample. T is specified by an ordered list of edge (Euclidean) lengths e_{ij} connecting certain pairs (x_i, x_j), $i \neq j$, along with a list of edge adjacency relations. The power weighted length of the tree T is the sum of all edge lengths raised to a power $\gamma \in (0, d)$, denoted by: $\sum_{e \in \mathrm{T}} |e|^\gamma$. The minimal spanning tree (MST) is the tree which has the minimal length $L(\mathrm{X}_n) = \min_{\mathrm{T}} \sum_{e \in \mathrm{T}} |e|^\gamma$. For any subset $\mathrm{X}_{n,k}$ of k points in X_n define $\mathrm{T}_{\mathrm{X}_{n,k}}$ the k-point MST which spans $\mathrm{X}_{n,k}$. The k-MST is defined as that k-point MST which has minimum length. Thus the k-MST spans a subset $\mathrm{X}^*_{n,k}$ defined by

$$L(\mathrm{X}^*_{n,k}) = \min_{\mathrm{X}_{n,k}} L(\mathrm{X}_{n,k})$$

The planar k-MST problem was shown to be NP-complete in [6]. Ravi *et al* proposed a greedy polynomial time algorithm for the planar k-MST problem with approximation ratio $O(k^{\frac{1}{4}})$.

Let $\nu \in (0, 1)$ be defined by $\nu = (d - \gamma)/d$ and define the statistic

$$\hat{H}_\nu(\mathrm{X}^*_{n,k}) = \frac{1}{1 - \nu} \ln \left(n^{-\nu} L(\mathrm{X}^*_{n,k}) \right) + \beta(\nu, d) \tag{4}$$

where β is a constant equal to the Rényi antropy of the uniform density on $[0, 1]^d$. In [5] Hero and Michel presented a d-dimensional extension of the planar k-MST approximation of Ravi et al, called the greedy k-MST approximation. In that paper we proved that when $k = \alpha n$, $\alpha \in [0, 1]$, and the length $L(\mathrm{X}^*_{n,k})$ of this approximation is substituted into (4) one obtains a strongly consistent and robust estimator of the Rényi entropy (1):

$$\hat{H}_\nu(\mathrm{X}^*_{n,k}) \to \min_{A : P(A) \geq \alpha} \frac{1}{1 - \nu} \ln \int_A f^\nu(x) dx \quad (a.s.)$$

where the minimization is performed over all d-dimensional Borel subsets of $[0, 1]^d$ having probability $P(A) = \int_A f(x) dx \geq \alpha$. This result was used in [4] to specify robust estimators of Rényi entropy which perform outlier rejection for the case that f is a mixture density of the form (3) with f_o uniform.

3. Extension: I-Divergence Estimation

Let $g(x)$ be a reference density on $\mathbf{R}^d$ which dominates the density $f(x)$ of a sample point $x =$

$[x^1, \ldots, x^d]^T$ in the sense that for all x such that $g(x) = 0$ we have $f(x) = 0$. For any x such that $g(x) > 0$ let $g(x)$ have the product representation $g(x) = g(x^1)g(x^2|x^1)\ldots g(x^d|x^{d-1}, \ldots, x^1)$ where $g(x^k|x^{k-1}, \ldots, x^1)$ denotes the conditional density associated with $g(x)$ of the k-th component. In what follows we will ignore the set $\{x : g(x) = 0\}$ since, as $f(x) = 0$ over this set, it has probability zero. Now consider generating the vector $y = [y^1, \ldots, y^d]^T \in \mathbf{R}^d$ by the following vector transformation

$$\begin{aligned} y^1 &= G(x^1) \\ y^2 &= G(x^2|x^1) \\ &\vdots \quad\quad \vdots \\ y^d &= G(x^d|x^{d-1}, \ldots, x^1) \end{aligned} \tag{5}$$

where $G(x^k|x^{k-1}, \ldots, x^1) = \int_{-\infty}^{x^k} g(\tilde{x}^k|x^{k-1}, \ldots, x^1)d\tilde{x}^k$ is the cumulative conditional distribution of the k-th component, which is monotone increasing except on the zero probability set $\{x : g(x) = 0\}$. Thus, except for this probability zero set, the conditional distribution has an inverse $x^k = G^{-1}(y^k|x^{k-1}, \ldots, x^1) = G^{-1}(y^k|y^{k-1}, \ldots, y^1)$ and it can be shown (via the standard Jacobian formula for transformation of variables) that the induced joint density, $h(y)$, of the vector y takes the form:

$$h(y) = \frac{f(G^{-1}(y^1), \ldots, G^{-1}(y^d|y^{d-1}, \ldots, y^1))}{g(G^{-1}(y^1), \ldots, G^{-1}(y^d|y^{d-1}, \ldots, y^1))} \tag{6}$$

Let $L(\mathrm{Y}^*_{n,k})$ denote the length of the greedy approximation to the k-MST constructed on the transformed random variables y, where $\mathrm{Y}^*_{n,k}$ is the set of k points spanned by this k-MST approximation. Then, from the results of [5] cited in the previous section, we know that

$$\hat{H}_\nu(\mathrm{Y}^*_{n,k}) \to \frac{1}{1-\nu} \ln \int h^\nu(y)dy \quad (a.s.) \tag{7}$$

Making the inverse transformation $y \to x$ specified by (5) in the above integral, noting that, by the Jacobian formula, $dy = g(x)dx$, and using the expression (6) for h, it easy to see that the integral in the right hand side of (7) is equivalent to the Rényi information divergence of $f(x)$ with respect to $g(x)$

$$\frac{1}{1-\nu} \ln \int h^\nu(y)dy = \frac{1}{1-\nu} \ln \int \left(\frac{f(x)}{g(x)}\right)^\nu g(x)dx.$$

Hence we have established that $\hat{H}_\nu(\mathrm{Y}^*_{n,k})$ is a strongly consistent estimator of the Rényi information divergence above. The results of [5] can thus be easily be extended to classification against any *arbitrary* distribution f_o, and not just the uniform distribution studied in [4].

4. Applications

100 samples were simulated from a triangle-uniform mixture density $f = (1-\epsilon)f_1 + \epsilon f_0$ where $f_1(x) = (\frac{1}{2} - |x^1 - \frac{1}{2}|)(\frac{1}{2} - |x^2 - \frac{1}{2}|)$ is a (separable) triangular shaped product density and $f_0 = 1$ is a uniform density, both supported on the unit square $x = (x^1, x^2) \in [0,1]^2$. The Rényi information divergences $I(f, f_0)$ and $I(f, f_1)$ were estimated by $\hat{H}_\nu(\mathrm{X}_n)$ and $\hat{H}_\nu(\mathrm{Y}_n)$, respectively. Y_n was obtained by applying the mapping $y = (y^1, y^2) = (F_1(x^1), F_1(x^2))$ to the data sample X_n, where $F_1(u)$ is the marginal cumulative distribution function associated with the triangular density.

In a first sequence of experiments the estimates $\hat{H}_\nu(\mathrm{X}_n)$ and $\hat{H}_\nu(\mathrm{X}_n)$ of the respective quantities $I(f, f_0)$ and $I(f, f_1)$ were thresholded to decide between the hypotheses $H_0 : \epsilon = 0$ vs. $H_1 : \epsilon \neq 0$ and $H_0 : \epsilon = 1$ vs. $H_1 : \epsilon \neq 1$, respectively. The receiver operating characteristic (ROC) curves are indicated in Figures **??** and **??**. Note that, as expected, in each case the detection performance improves as the difference between the assumed H_0 and H_1 densities increases.

In a second sequence of experiments we selected two realizations of the triangle-uniform mixture model for the values $\epsilon = 0.1$ and $\epsilon = 0.9$. For the former case the triangular is the dominating density and for the latter case the uniform is the dominating density. In each case the k-MST was implemented ($k = 90$) as a robust clustering algorithm to identify data points from the dominating densities - in the former case the k-MST was applied directly to X_n while in the latter case it was applied to Y_n. The resulting k-MST quantities $\hat{H}_\nu(\mathrm{X}_{n,k})$ and $\hat{H}_\nu(\mathrm{Y}_{n,k})$ can be interpreted as robust estimates of the uncontaminated Rényi information divergences $I(f_1, f_0)$ and $I(f_0, f_1)$. respectively. Figure **??** and **??** illustrate the effectiveness of these estimates as "outlier rejection" algorithms.

References

[1] M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, pp. 349–369, 1989.

[2] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. van der Meulen, "Nonparametric entropy estimation: an

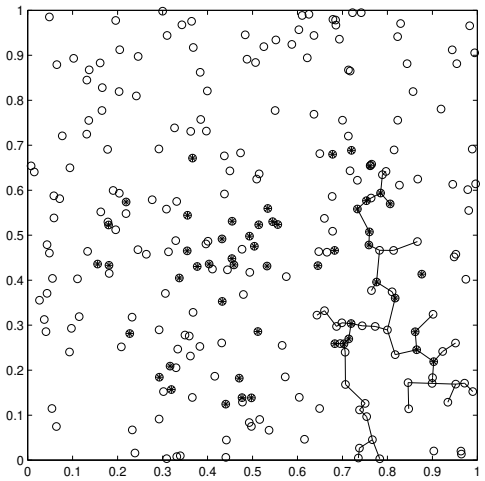

Figure 1. *ROC curves for the Rényi information divergence test for detecting triangle-uniform mixture density $f = (1-\epsilon)f_1 + \epsilon f_0$ (H_1) against the uniform hypothesis $f = f_0$ (H_0). Curves are indexed by values of ϵ ranging from 0.1 to 0.9 in equal steps.*

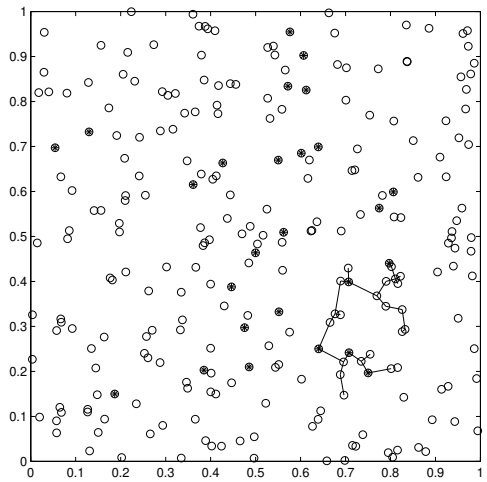

Figure 3. *A scatterplot of a 100 point sample from triangle-uniform mixture density with $\epsilon = 0.1$. Labels 'o' and '*' mark those realizations from the uniform and triangular densities, respectively. Superimposed is the k-MST implemented directly on the scatterplot X_n with $k = 80$.*

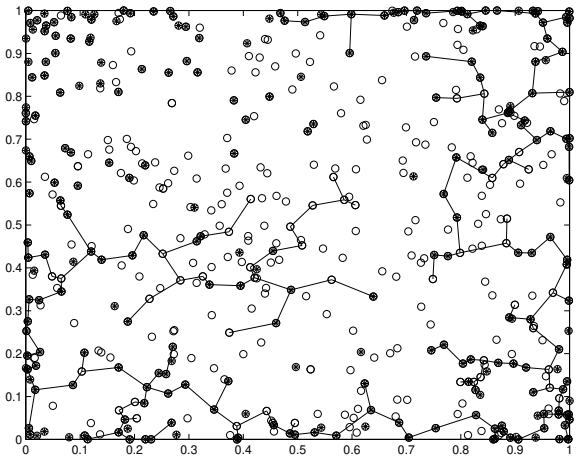

Figure 2. *ROC curves for the Rényi information divergence test for detecting triangle-uniform mixture density $f = (1-\epsilon)f_1 + \epsilon f_0$ (H_1) against the triangular hypothesis $f = f_1$ (H_0). Curves are indexed by values of ϵ ranging from 0.1 to 0.9 in equal steps.*

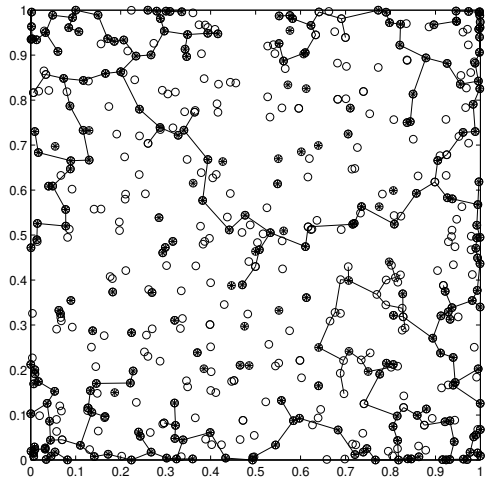

Figure 4. *A scatterplot of a 100 point sample from triangle-uniform mixture density with $\epsilon = 0.9$. Labels 'o' and '*' mark those realizations from the triangular and uniform densities, respectively. Superimposed is the k-MST implemented on the transformed scatterplot Y_n with $k = 80$.*

overview," *Intern. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.

[3] I. Csiszár, "Information-type measures of divergence of probability distributions and indirect observations," *Studia Sci. Math. Hung.*, vol. 2, pp. 299–318, 1967.

[4] A. Hero and O. Michel, "Robust entropy estimation strategies based on edge weighted random graphs," in *Proc. of Meeting of Intl. Soc. for Optical Engin. (SPIE)*, San Diego, CA, July 1998.

[5] A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. on Inform. Theory*, vol. to appear, , 1999.

[6] R. Ravi, M. Marathe, D. Rosenkrantz, and S. Ravi, "Spanning trees short or small," in *Proc. 5th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 546–555, Arlington, VA, 1994.

[7] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, volume 1, pp. 547–561, 1961.

[8] P. Viola and W. Wells, "Alignment by maximization of mutual information," in *Proc. of 5th Int. Conf. on Computer Vision, MIT*, volume 1, pp. 16–23, 1995.